



Teaching an LLM to Catch British Sarcasm

Sentiment detection on landlord calls in production

● **The problem: quiet churn**

- Managed thousands of rental properties — retention drives the economics.
- A big part of churn is unpreventable (landlords sell or move back in) — but not most of it.
- The rest — switching to a competitor or self-managing — is service-driven, preventable, and expensive.
- Need: catch it early, hidden inside tens of thousands of mostly-polite messages.

● **Why off-the-shelf sentiment fails**

- Standard models assume emotion words track the feeling. British complaint removes them.
- “A little disappointed” · “when you get a chance” · “I’d have thought...” — all score positive.
- NPS misses it too: the quiet leavers are too polite to give a low score.
- And much dissatisfaction is our silence — unanswered emails, missed callbacks. No message to classify.

● The data

11,336

interactions

~2,000

unique landlords

4

channels (email & calls, CRM & not)

~1,100

calls transcribed (Voximplant)

Plus hand-labelled ~200 borderline cases — explicit decision rules, not judgement calls.

● **The classifier: two-step pipeline**

- Step 1 — is this negative at all? (positive / neutral / negative).
- Step 2 — runs only on the ~6–7% negatives: severity, topic, intent, routing.
- Splitting the steps tunes detection recall separately from classification precision.
- LLM-based (Claude Sonnet), scoring continuously off BigQuery. ~\$0.003 per message.

● **Where the work went: prompt engineering**

- Stop rewarding politeness — classify on the substance under the courtesy.
- Stop over-escalating — make severity a checklist (orange = two signals, red = explicit exit language).
- Separate venting from blame — negative tone aimed elsewhere ≠ service failure.
- Hand-labelled polite-but-negative cases used as few-shot anchors.

● What we found

- Calls run ~50% more negative than email — people are guarded in writing, direct on the phone.
- Channel switching (email → phone) is a leading indicator of escalation.
- 36% of all negativity = communication failure (can't reach a human).
- Competitor-switchers run far higher negativity — expressed calmly, not angrily.
- Coverage bounds quality: feed every channel, or the picture is confidently wrong.

● **From detection to action**

- Every intent maps to a defined action, owner and SLA — agreed in advance.
- Communication failure → manager callback in 24h. Self-management signal → personal value call.
- The 36% finding pointed to a structural fix, not more chasing...
- ... which led to a second build: an agentic voice assistant for inbound calls.

● What generalises

- Don't use off-the-shelf sentiment — model how your customers actually talk.
- Two-step pipeline: a cheap, sensitive detector, then structured classification only on what it flags.
- Invest in labelling and prompts, not the model — explicit rules, hard cases as few-shot anchors.
- Detection needs an owner and an SLA — otherwise it's just a dashboard.
- Coverage bounds quality — feed every channel, or the output is confidently wrong.